View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Procedia Computer Science

Volume 29, 2014, Pages 2462–2471

ICCS 2014. 14th International Conference on Computational Science

Social networks mining for analysis and modeling drugs usage

Andrei Yakushev[1] and Sergey Mityagin[1]

[1]*ITMO University, Saint-Petersburg, Russia.*

andrew.yakushev@ya.ru, mityagin@iac.spb.ru

Abstract

This paper presents approach for mining and analysis of data from social media which is based on using Map Reduce model for processing big amounts of data and on using composite applications for performing more sophisticated analysis which are executed on environment for distributed computing-based cloud platform. We applied this system for creation characteristics of users who write about drugs and to estimate factors that can be used as part of model for prediction drug usage level in real world. We propose to use social media as an additional data source which complement official data sources for analysis and modeling illegal activities in society.

Keywords: social media, data mining, gig data, illicit drug use, map reduce, clavire, feature selection

1 Introduction

The growth of modern society, among others, has led to increasing role of information development and using of information technology in everyday life. Nowadays, the speed of information and the ease of access to it are one of the criterions for the society development. In this situation the prominent role is played by the social media which provides an opportunity to share information about some topic among interested users.

Social media refers to a group of Internet-based applications that allows users to create and exchange content [1]. Examples of social media are Social Network Sites (SNS), blogs and microblogs, collaborative projects and etc. Openness of some social networks allows mining data which should be stored in local database and lately can be used to study processes that take place in social media. Social media usually considered as a kind of real world reflection and sometime used for analysis of real society. Since computerization has touched all without exception spheres of society, it is of interest to assess the impact of the development of information technologies and especially social media to the illegal activities in society.

This paper describes our solution for mining and analysis data from social media. Our approach for data processing consist from three stages and combines the use of two different concepts – big data and cloud computing. First and second stages are data mining from social media and its filtration or

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014
© The Authors. Published by Elsevier B.V.

doi:10.1016/j.procs.2014.05.230

aggregation which in result gives relatively small datasets with data relevant to the solving task. Data mining is performed by crawler which is based in MapReduce model for distributed computations and which we implemented using Hadoop framework [2]. On the last stage obtained small datasets are analyzed using sophisticated models. Whole data analysis process is formalized in composite application which is run in our environment for distributed computing-based cloud platform CLAVIRE (CLoud Applications VIRtual Environment) [3].

Another part of the paper describes analysis of people who write about drugs in social media. We present an idea of using social media as an additional data source for analysis and modeling of illegal activities in society. Developed technologies for mining and analysis are applied to characterize users who write about drugs. Characteristics reveal additional interests of users and compose their psychological portrait. This paper also describes prediction model for the level of drug use among population which considers various factors, like macro-state of the population and individual characteristics of residents. Results of social media analysis such as level of interest to the drug theme or characteristics of users who uses drugs can be used to increase accuracy of this model.

Despite the relatively recent appearance of the term Big Data in 2008 [4] this area of science attracted huge attention of business and academia. Many tasks require analytics of huge amounts of data, for example, experiments on Large Hadron Collider [5], climate simulations [6] and creation a recommendation systems for internet services [7], but the need of analytical solutions is especially important for the business. The cost-effective solution for business is to use clouds systems which will provide transparent resource and task management [8]. Unfortunately data management and its analytics in the clouds give a lot of challenges [9] and currently there is no best way to overpass them [10].

Social media are increasingly being used in academia and business to analyze the different processes of the real world, for example stock market prediction [11], viral marketing [12], correlation with the incidence of influenza [13] and identifying potential new harmful drug side effects [14]. There are also plenty of researches concerning drug addiction, for example, about connection between needle-sharing patterns and real world social networks [15] and social networks influence on the transition to injecting drug use [16] which studies people physically addicted to drugs and their friendship network. This paper is the first that tries to use data from social media i.e. virtual world to study drug usage. Main advantage of this work is that it provides insights on hardly observable group of people who rarely use drugs or do not have addiction to them.

Section 2 of this paper describes background information about drug addiction, defines groups of people formed by the level of drug addiction, and describes data sources for analysis of drug addicted people. Section 3 describes our approach of data analysis which is based on combining big data with cloud platform. Section 4 describes analysis of people who writes about drugs in blog platform Livejournal. Section 5 describes model for prediction level of drug consumption.

2 Background

Research of the drug usage touches many questions, some of which are addressed in this paper: measurement and prediction of the drug consumption level among population, drawing up a characteristic of people who uses drugs. But many other questions which are rather a pure sociology could be asked: what are the reasons of a person to use drugs and how to reduce level of illicit drug usage. To get a chance to answer these questions many aspects of the drug usage should be analyzed.

Like many social processes, drug usage analysis gives a lot challenges because many hardly formalized factors should be taken into account. For example different types of drugs have different

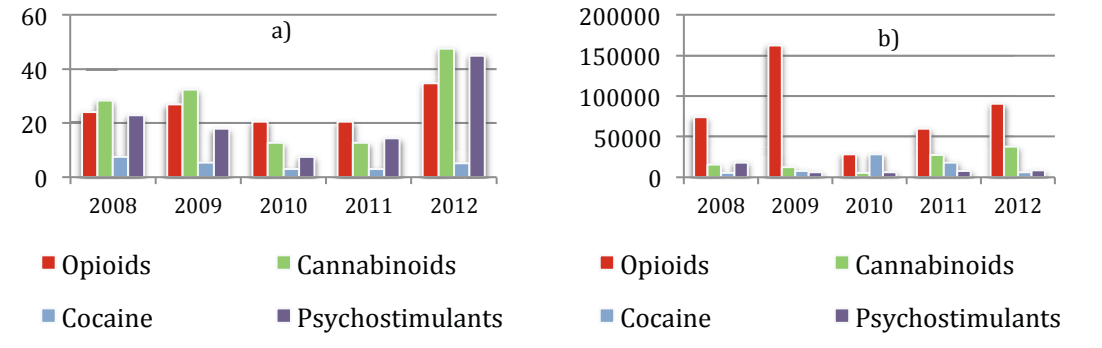

Figure 1. a) Prevalence of the drug type as a percent of respondents confirmed their experience of drug usage. b) Statistics about the weight of seized drugs of different types.

levels of consumptions among different age and sex groups. Also cultural, subcultural and ethnographic aspects of the population and economic level of the area have influence on the drug usage [17]. Moreover level of the drug usage changes over the time because of the changes in many factors: social censure which encourages the consumption of any drug, law enforcement activities, availability of advocacy and other changes in a particular culture.

It is important to divide population of the territory into groups based on the drug usage level. This division changes over the time because individual can change level of the drug usage and transfer to another group. Usually distinguish the following groups of people [17] which are hereinafter referred by Roman numerals: not related to drug usage and not susceptible to their consumption (I); having prerequisites to drug usage, as they consist in the risk group (II) possibly connected with interests, age, social group, etc.; drug users who do not have forms of dependence (III); consumers having a psychological dependence on drugs (IV); consumers having a physical dependence on the drug (V). It should be noted that it is possible for a man to have both psychological and physical dependences and in this case we will refer him to V group. Modeling of the drug usage denote to prediction of the sizes of the defined groups or how people transfer from one of these groups to another.

It is also important to consider different drug types separately because they differently influence on how individual gets psychological or physical dependences on the drug use. Following the classification of drugs, adopted the Law on controlled Psychotropic Substances (Controlled Substances Act) (CSA) all drugs fall into five categories [17]: psychedelic drugs or hallucinogens, stimulants, opiates, depressants or sedative-hypnotic drugs. Categories of danger of physical dependence are next: drugs of high danger - opiates, some hallucinogens (heroin, LSD); moderate hazard drugs - hallucinogens, stimulants, depressants (codeine, steroids, barbiturates); low hazard drugs. Obviously, the use of drugs causing physical dependence makes man adjust their activity in connection with the need to receive regular doses of drugs, search and spend money on a drug. Also drug usage can change person lifestyle so his new environment will also share his interests.

Every drug type has target group of people with specific lifestyle. Opiates are popular among people who have physical and psychological dependence and who got strong drug addiction. And psychedelic drugs are popular among people who are interested in expansion and altering of the human mind. The question of finding these target groups for different drug types is quite interesting and still unresolved. Answering on this question can help in creating anti-drug propaganda which will reduce their using.

In Russia analysis of drugs usage is quite challenging because of its illegal and secrecy nature. Only few sources of information exist which, unfortunately, are biased and are not complete. Direct interviews about drug use experience are possible but sample size is small and data is biased to the soft drugs (see Fig. 1.a). It is also possible to track number of people who have physical dependence

on the drugs and who have been "registered" by narcologists. This data source can give information only about groups IV - V and is strongly biased to people who use opium or heroin. Soft drugs are not at all represented in this data. The most valuable source of information is official statistics about seizure of illegal drugs (see Fig 1.b) since it shows state of the drug market which in turn correlates with demand and usage of the drugs among the population.

We propose to use social media as an additional source of information about drugs usage. First of all they can give information on the level of interests to the drugs topics and secondly they can provide information about different subcultural and psychological characteristics of people who uses drugs. This information can be used as additional factors to model level of the drug usage and to create more efficient anti-drug advertisements.

3 Data Management

This section describes our management of data from social media which provides unified approach for solving scientific tasks. Because social media contains huge amount of data we used BigData paradigms to mine and analyze it. Firstly, data from social media is mined using our crawler [2] which stores it into Hadoop cluster. Secondly, big volume of mined data is filtered and aggregated in order to get relatively small datasets of information that is relevant to the solving task. Finally aggregated data is used as an input for composite applications which perform final and sophisticated data analysis. To organize computational process of the composite application we used AaaS (Application as a Service) model which is implemented in our environment for distributed computing-based cloud platform CLAVIRE (CLoud Applications VIRtual Environment) [3]. Composite application operates with data which is obtained from Hadoop cluster using developed API and calculations are performed by computational module which can be seamlessly integrated in CLAVIRE. As a result this provides unified approach for mining and analysis data from social media.

Data from social media is obtained from the API (Application Programming Interface) provided by social media sites. Social media contains different types of data – user information, connections between users, generated by users' content and etc. Each data type usually accessed by separate API and each API impose restrictions to the amount of accesses. This lead to the impossibility to collect in reasonable time all data stored in social media. But even with these restrictions amounts of data are big. Also data types can be structured (e.g. user profiles, links between users) and unstructured (e.g. user's interests, posts or pictures). To efficiently work with big and unstructured data we use Hadoop framework which implements MapReduce model for distributed computations and other associated with it technologies. Another important characteristic of the data from social media is its «sparseness» which lies in the fact that only small amount of data is relevant to the solving problem. This means that original big data mined from social media should be filtered, mapped or aggregated to some small dataset that is actually used in further analysis.

Hadoop framework consists from two main components: the distributed file system HDFS and component that performs data processing in MapReduce model. HDFS uses NoSQL data model in which records consist of key and value parts of arbitrary type. Records forms table which on physical level are stored by parts (chunks) on several machines that works under Hadoop framework and processed in place. HDFS is highly scalable, reliable and at the same time flexible. One of its benefits is ability to store data on physical level in different formats. We used Google Protobuf binary serialization mechanism to store retrieved from social networks data. Protobuf also simplifies handling situations with data format changing. To increase performance of the HDFS we compress data using Snappy compression format which is designed to fast compress and decompress data but have average compress ratio.

In the problems we faced, different data types were aggregated separately but further aggregated small datasets were considered together in order to get final result. In turn final analysis required the

work of several steps each representing some basic algorithm or analysis. This inspired us to formalize this analysis as a workflows or composite applications which are run on CLAVIRE. CLAVIRE is a distributed computing-based cloud platform of the second-generation that implements paradigm AaaS (Application as a Service) and which provided seamless integration of separate software module.

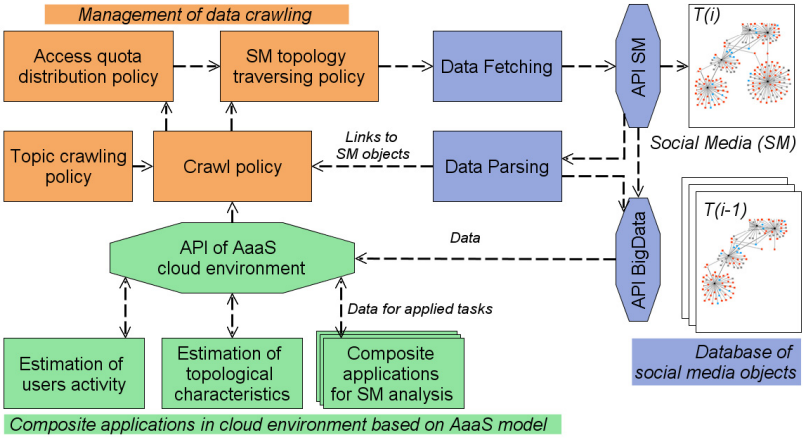

Figure 2: Iterative crawling procedure.

Our approach is illustrated with two examples. First one describes iterative algorithm for mining data from social media (see Fig. 2). On each iteration set of requests to social media is processed and responses are written to HDFS. This data can be analyzed directly on the Hadoop cluster for example to estimate activity of users or to obtain topological characteristics of the networks. But data can be also provided to some composite application that performs more complex data analysis. This application is run in distributed computational environment CLAVIRE. Results of this analysis are used to control crawling process and to create lists of requests that will be processed on the next iteration. Crawler control procedure is also implemented as a composite application which solves three subtasks: filtration of objects relevant to some topic, quota distribution between already crawled object of social media and choosing which new objects should be crawled on the next iteration.

Another example (see Fig. 3) describes application that is used to find communities of users that are interested in same topics. On the first step data from social network is mined, and then data about users' interests and tags that they used to mark posts are extracted. These steps are performed using Hadoop cluster. At the same time friendship network is analyzed and community structure is extracted. And finally for each community the most frequent interest and tag sets are extracted. These steps are realized as standalone applications which can be reused for many different networks. Union of these steps in workflow forms composite application which can be further transparently used by clients of the system.

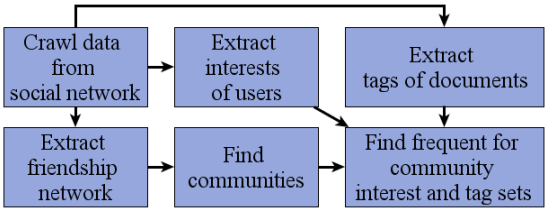

Figure 3: Examples of workflows that run on distributed computing-based cloud platform.

4 Data Analytics

This section describes analysis of users that have blogs at Livejournal SNS and which can be used in model of drug usage in society. We analyzed users who writes about drugs in their blogs and created their characteristics by analyzing interests (Dijkstra & Yakushev, 2012). Characteristics consist of set of topics of interest that arise this group of users significantly more (or less) often. This allows to create psychological portrait which describes features that distinguish members of this group from all other users and to determine their lifestyle. Described in this section approach can be also used to create characteristics of any group of users and it has practical importance in case of analyzing group of users with common feature which is rather difficult to obtain in real world.

It should be noted that sets of people who use drugs in real world and who write about them in social media are not the same, however they should overlap. User may have many reasons to write about drugs: cure drug addiction may be his profession; he may try to draw attention to the problem of increased level of drug usage; user may write about not using drugs and healthy lifestyle; user may simply use drug slang as a joke. But we believe that in the huge amount of data from social media contains information about users that share their experience of drug usage. And by analyzing users who writes about drugs it is possible to extract information about people who uses drugs.

To identify users who write about drugs we used dictionary of the official and slang keywords and keyphrases. Each keyword was assigned to one of the nine groups depending on the semantic meaning of the word. We identified following groups of words: heroin, marijuana, cocaine, raw opium, pills, injections, preparation slang, syringe, needle, methamphetamine, ephedrine and general words. Keyphrases were defined based on the predefined rules which in turn were based on the semantical groups of keywords. For example we used following rules to automatically create set of keyphrases: “drugs” + “ways of using”; “drugs” + “preparation”; “drugs” + “effects” and “drug” + “drug synonym”. We also did manual revision of the keyphrases in order to add new phrases or to remove non-relevant to drug theme phrases. We also manually split keywords on the three groups depending on their “strength” of belonging to the drug theme. Each group has same weight which is used to estimate relevance of the text to the drug theme. We made keyphrase weight larger than a sum of its keywords weight to indicate that keyphrase is stronger signal of document relevance to the drug theme. Weight of the document is calculated as a sum of all keywords and keyphrases that are found in it. Top 20% documents with largest weights are assumed to be relevant to the drug theme. If user has at least one document relevant to the drug theme then he is assumed to be interested in drug theme.

Our dictionary consists from 368 keywords and 8359 keyphrases. We mined data for about 100000 randomly selected users from Livejournal Social Network Site (SNS). For each user we know his last 25 posts, his in-coming and out-going connections with other users, data that he specified in his profile (real name, age, place of living) and his interests – keywords that describes user’s areas of interests (e.g. music, films, computers or sport). Using the method described above we identified 16553 users who writes about drugs. We will further call this group of users as drug community.

Let $I = \{I_1, \dots, I_m\}$ be set of interests specified by all users in data set. Each interest is described by set of users that defined it in their profiles and for each user it is known whether he is from drug community or not. Then for each interest I_i should be determined whether it appears significantly more or less often in drug community. Let $S = \{S_1, \dots, S_m\}$ be this set of significant interests based on which we will draw up characteristic of users from drug community.

To check if interest is significant or not we use two-tailed Fisher Exact Test for 2×2 contingency tables (Fisher, 1922). Each interest is described by four values: total number of users from drug community specified it in their profiles, total number users from drug community that do not specified it in their profiles and similar two values for users that are not from drug community. Based on these four values 2×2 contingency table is created for which Fisher Exact Test is applied. Fisher Exact Test check hypothesis H_i^0 that interest is significant and returns for each interest I_i a corresponding p-value

p_i. By comparing p-value p_i with the fixed significance level p it is possible to determine set of significant interests S.

Set S may contain interests mistakenly called significant (false discoveries) despite using threshold for significance level p for each interest. This may happen due the large number of interests for each of which one hypothesis is checked. To upper-bound percentage of false discoveries by value q in the set S we used Benjamini and Hochberg controlling procedure for multiple hypotheses testing (Benjamini & Yekutieli, 2001) which rejects several interests. Finally set of significant interests S is formed.

Set of significant interests S used to create characteristics or psychological portrait of users. To simplify manual analysis of this set we automatically grouped interests in more general topics. For this purpose we used hierarchical agglomerative clustering algorithm with a complete linkage strategy (Everitt, Landau, & Leese, 2001) with similarity between two interests defined as: $sim(S_1, S_2) = \frac{n(U_1 \cap U_2)}{\sqrt{n(U_1) \times n(U_2)}}$, where U_1 and U_2 are sets of users that mentioned in their profiles interests S_1 and S_2 respectively. Described similarity measure is also known as cosine similarity.

Statistical analysis of the dataset reveals 268 significant interests of the 3282 interests which appeared more than 10 times in dataset. Clustering of the interests gave 42 different themes. Names of themes compiled by the authors and thus have some subjective. Figure 4 shows the most popular and unpopular themes in drug community and rest of the dataset. For each theme we calculated probability of its occurrence in drug community and rest of the dataset as a probability of occurrence at least of one interest from that theme.

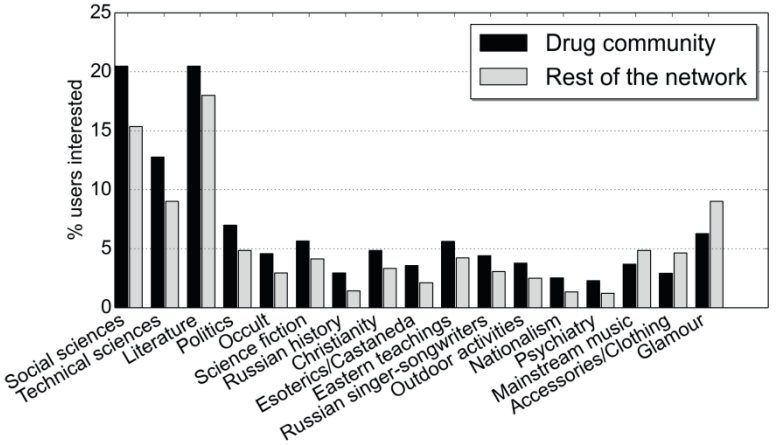

Figure 4: Characteristic of users writing about drugs. For each theme its popularity in drug community and rest of the dataset is specified.

5 Modelling

This section describes modeling of drug usage among population of the specific area. Modeling involves determining and prediction of sizes of different groups of people that uses drugs (e.g. groups I, II, III, IV and V described in Section 2) and is based on the analysis of different factors that influence on the drug usage. Modeling should take into account population dynamics on the territory and the possibility of individual to become a drug addict and the possibility of recovery. As factors that influence on the drug usage can be used macro and micro characteristics of the territory and society. It is important to note that influence of the factor differs depending on the age and sex of individuals. Among the most important groups of factors are the following:

Φ_1 - Socioeconomic factors determining the level of life of the Territory as a whole, including the characteristics of the stratification of society, improvement, social security and other factors. Dynamic factors can be estimated from the observed data and official statistical indicators provided by government. However, since official statistical reports are used, significant latency of observed values should be taken into account. In general group of factors Φ_1 characterizes the overall «condition» of the territory.

Φ_2 - Factors of emotional and psychological state of the society that are indicators of well-being. These factors are usually determined by sociological researches and are not always objective because, for example, highly susceptible to propaganda.

Φ_3 - Affiliation of an individual to some classes of society. This factor is reasonable because sociological researches (LINK [1]) shown that some classes of society have greater susceptibility to drug use.

Φ_4 - Individual has interests similar to people who uses drugs. Presence of drug-related interests can signal that user belongs to some subculture or communities that endorse drug use of any type. Information about individual interests and common interests of people who uses drugs can be extracted from social media.

Model divides each group $k \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}, \mathrm{V}\}$ on subgroups depending on age and sex of people. Denote size of each subgroup as n_{ij}^k where i is age and j is sex of people in subgroup. The goal is to express n_{ij}^k as a function of factors that influence on the level of drug usage in group k: $n_{ij}^k = F(\Phi_1, \Phi_2, \Phi_3, \Phi_4)$. Shape and parameters of the functional F can be adjusted based on the retrospective statistical data, sociological studies and relevant data of social networks and the public. Examples of values of some factors are presented in table 1.

Age group	Unemployment level	Gini index	Ratio of mortality and fertility rates	Ratio of divorces and marriages	Life satisfaction
15-17	0.3692	-0.7886	0.7539	0.6892	0.6855
18-19	0.5143	-0.8771	0.8469	0.8501	0.6516
20-39	-0.3822	0.8188	-0.9837	-0.8513	-0.8438
40-59	-0.2032	0.6967	-0.7679	-0.8169	-0.8951

Table 1. Pierson correlation between age and factors for group of people with physical drug addiction.

The state of all age and sex subgroups in group k can be described by matrix N^k whose general view is presented at (1):

$$N^k = \begin{pmatrix} n_{1,1}^k & n_{2,1}^k \\ n_{1,2}^k & n_{2,2}^k \\ \vdots & \vdots \\ n_{1,100}^k & n_{2,100}^k \end{pmatrix} \quad (1) \qquad A = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \hat{f}_i^1 & 0 & \dots & 0 & 0 \\ 0 & \hat{f}_i^2 & \dots & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & \hat{f}_i^{n-1} & 0 \end{pmatrix} \quad (2)$$

As noted earlier, each group characterized by matrix N^k changes over time. Evolution of N^k can be described in terms of Markov chain by matrix equation: $N_{t+1}^k = AN_t^k + W_{k-1}$, where A – matrix of probabilities of the individual with corresponding sex and age fall within one of five groups I-V, and W_k - net migration from the adjacent group to the group k. General view of the matrix A is presented at (2). Last equation allows to turn from estimations of sizes of groups n_{ij}^k to estimations of transitions between groups $\hat{f}_{ij}^k$ which are also can be used as a basis for the construction of forecasts of drug use in the territory.

6 Conclusion and Discussion

In this paper we presented our approach for mining and analysis data from social media which is based on combination big data and cloud computing paradigms. Map Reduce model is used to mine, store and process big amounts of data from social media. Using Hadoop framework we implemented distributed topic crawler for social media. Processing of mined data is also performed by Hadoop which simplifies development of new algorithms and provides high scalability and flexibility.

We proposed to use cloud computing environment for distributed computations which simplifies creation and execution of composite applications which perform complex data analysis. Composite application operates with data accessed from Hadoop by API and consists of atomic computational tasks that achieve some analysis and modeling goals. We used environment for distributed computing-based cloud platform CLAVIRE which implements AaaS model for cloud computing and provide seamless integration of separate software modules and that runs on different computing resources. As a result this strongly simplified designing, executing and reusing already implemented applications.

Social media contain a lot of personal information and can be used as an additional data source for analysis of social processes in real world. Especially processes that are hardly observable in real world like spread of infections or interest and attitude of society to drug usage. Despite the fact that social media cannot replace traditional sources of information are available from social researches or official statistical indicators provided by government, they can complement this data. Another advantage of social media is that they provide macro and micro characteristics of users. Macro characteristics are general interest of whole population of users to some topic and micro characteristics are for example centrality of user to some community. And again social media reveal knowledge about micro level of social processes while other sources of information are more about macro parameters.

We proposed to use data from social media as an additional source of information for analysis and modeling of drug usage among population of certain territory. Analysis at the micro-level of users who write about drugs gave insight about their interests which distinguish them from others. We obtained their psychological portrait which consists of topics of interest that appear more (or less) often in this group. We used dictionary of drug related keywords and key phrases to determine if user writes in blog posts about drugs. We think that among different users who have different reasons to write about drugs we found some set of users who share their experience of using drugs. This is indirectly confirmed by the presence of such topics as Russian rock, non-traditional medicine, occult, eastern teachings, and esotericism which in Russia are generally considered drug-related. But to confirm this further research is required. Described approach uses many different parameters and thresholds which significantly influence on the list of significant interests. However our experiments showed that list of significant topics remains quite stable.

We described model for prediction of number of people that have different levels of drug addiction on the specified area. Prediction is based on factors of the macro and micro state of the society on the specified territory. Several macro factors can be estimated based on the official data sources, but level of interest to drugs is much easier to estimate based on the data from social media. Moreover official data sources cover people with strong physical drug addiction but social media cover people with light drug addiction forms. We showed estimations of factors values for strong addicted people but estimation of factors based on the data from social media as well as modeling with these factors is a future work.

However some criticism about using social media should be noted, for example not all society groups are presented equally in social media and their main consumers are young people. Also there are criticisms about trustworthiness [22] and reliability [23] of using data from social media which look quite reasonable. Despite these criticisms, we believe that the study of social networks is important and can reveal information about processes that are hidden or can’t be directly observed in society. This work was financially supported by the Government of the Russian Federation, Grant 074-U01.

References

[1] M. Kaplan, M. Haenlein, Users of the world, unite! The challenges and opportunities of Social Media, *Bus. Horiz.*, 53:59–68, 2010.

[2] A. Yakushev, A. Boukhanovsky, P.A. Sloot, Topic Crawler for Social Networks Monitoring, *P. Klinov, D. Mouromtsev (Eds.), Knowl. Eng. Semant. Web SE - 17, Springer Berlin Heidelberg*, 214–227, 2013.

[3] K. V Knyazkov, S. V Kovalchuk, T. N Tchurov, S. V Maryin, A. V Boukhanovsky, CLAVIRE: e-Science infrastructure for data-driven computing, *J. Comput. Sci.*, 3:504–510, 2012.

[4] Community cleverness required, *Nature*, 455:1, 2008.

[5] C. Lynch, Big data: How do your data grow?, *Nature*, 455:28–29, 2008.

[6] Jenny Mangeksdorf, Supercomputing the Climate: NASA's Big Data Mission, *Comput. Inf. Sci. Technol. Off.*, 2012.

[7] V. Borkar, M.J. Carey, C. Li, Inside Big Data management: ogres, onions, or parfaits?, *Proc. 15th Int. Conf. Extending Database Technol.*, 3–14, 2012.

[8] X. Sun, B. Gao, Y. Zhang, W. An, H. Cao, C. Guo, et al., Towards Delivering Analytical Solutions in Cloud: Business Models and Technical Challenges, *E-Bus. Eng. (ICEBE), 2011 IEEE 8th Int. Conf.*, 347–351, 2011.

[9] D.J. Abadi, Data Management in the Cloud: Limitations and Opportunities., *IEEE Data Eng. Bull.*, 32:3–12, 2009.

[10] M.D. Assun, Big Data Computing and Clouds : Challenges, Solutions, and Future Directions, *arXiv preprint arXive:1312.4722*, 2013.

[11] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, *J. Comput. Sci.*, 2:1–8, 2011.

[12] P. Domingos, Mining social networks for viral marketing, *IEEE Intell. Syst.*, 20:80–82, 2005.

[13] A.F. Dugas, Y.-H. Hsieh, S.R. Levin, J.M. Pines, D.P. Mareiniss, A. Mohareb, et al., Google Flu Trends: correlation with emergency department influenza rates and crowding metrics., *Clin. Infect. Dis.*, 54:463–469, 2012.

[14] David Shaywitz and Mathai Mammen, The next killer app, *Boston Globe*, 2011.

[15] J.B. Unger, M.D. Kipke, C.J. De Rosa, J. Hyde, A. Ritt-Olson, S. Montgomery, Needle-sharing among young IV drug users and their social network members: The influence of the injection partner's characteristics on HIV risk behavior, *Addict. Behav.*, 31:1607–1618, 2006.

[16] A. Neaigus, V.A. Gyarmathy, M. Miller, V.M. Frajzyngier, S.R. Friedman, D.C. Des Jarlais, Transitions to injecting drug use among noninjecting heroin users: social network influence and individual susceptibility, *JAIDS J. Acquir. Immune Defic. Syndr.*, 41:493–503, 2006.

[17] C.R. Bartol, A.M. Bartol, Criminal behavior: A psychosocial approach, *Upper Saddle River, NJ: Prentice Hall*, 1999.

[18] L. Dijkstra, A. Yakushev, Inference of the Russian drug community from one of the largest social networks in the Russian Federation, *Quality & Quantity*, 2013.

[19] R. Fisher, On the interpretation of X2 from contingency tables, and the calculation of P, *J. R. Stat. Soc.*, 85:87–94, 1922.

[20] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *Ann. Stat.*, 29:1165–1188, 2001.

[21] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, *Arnold*, 2001.

[22] A. Kittur, B. Suh, E.H. Chi, Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia, *Proc. 2008 ACM Conf. Comput. Support. Coop. Work*, 477–480, 2008.

[23] M.R. Auer, The policy sciences of social media, *Policy Stud. J.*, 39:709–736, 2011.